Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding

Song Han¹, Huizi Mao², William J. Dally¹,³
¹ Stanford University  ² Tsinghua University  ³ NVIDIA
{sghan, dally}@stanford.edu  mhz12@mails.tsinghua.edu.cn

Deep Compression is a three stage compression pipeline: pruning, quantization and Huffman coding. Pruning reduces the number of weights by 10x, quantization further improves the compression rate between 27x and 31x. Huffman coding gives more compression: between 35x and 49x. The compression rate already included the metadata for sparse representation. Deep Compression doesn’t incur loss of accuracy.

Motivation: Make DNN Smaller

Pruning, Relative Indexing & Huffman Coding

We pruned, quantized, and Huffman encoded four networks: Lenet-5, Lenet-300-100 on MNIST and AlexNet, VGG-16 on ImageNet. The compression pipeline reduces network storage by 35x to 49x across different networks without loss of accuracy. The total size of AlexNet decreased from 240MB to 6.9MB, which is small enough to be put into on-chip SRAM, eliminating the need to store the model in energy-consuming DRAM memory.

Results

Accuracy v.s. compression rate under different compression methods. Pruning and quantization works best when combined.

Table 1: Compression statistics for AlexNet. P: pruning, Q: quantization, H:Huffman coding.

Table 2: Compression statistics for VGG-16. P: pruning, Q: quantization, H:Huffman coding.